Critical Vector Learning for Text Categorisation

Lei Zhang, Debbie Zhang, and Simeon J. Simoff

Faculty of Information Technology, University of Technology, Sydney
PO Box 123 Broadway NSW 2007 Australia
{leizhang, debbiez, simeon}@it.uts.edu.au

Abstract. This paper proposes a new text categorisation method based on the critical vector learning algorithm. By using the proposed approach, the number of support vectors has been significantly reduced by implementing a Bayesian treatment of a generalised linear model with identical function form to the function form of support vector machines approach. This leads to much reduced computational complexity in the prediction process, which is critical in online applications.

Key words: Support Vector Machine, Relevance Vector Machine, Critical Vector Learning, Text Classification

1 Introduction

Text categorisation is the classification of natural text or hypertext documents into a fixed number of predefined categories based on their content. Many machine learning approaches have been used in the text classification problem [1]. One of the leading approaches is the support vector machine (SVM) [2], which achieved remarkable performance in many applications, particular in text categorisation [3–6]. SVM has good performance on large data sets and scales well. It is linear efficient and scalable to large document sets. Using the Reuters News Data Sets, Rennie [7] compared the SVM with Nave Bayes algorithm based on two data sets: 19,997 news related documents in 20 categories and 9649 industry sector data documents in 105 categories. Another researcher Joachims [8] compared the performance of several algorithms with SVM by using 12,902 documents from the Reuters 21578 document set and 20,000 medical abstracts from the Ohsumed corpus. Both Rennie and Joachims has shown that SVM performed better.

Tipping [9] introduced the relevance vector machine (RVM) methods. These methods can be viewed from a Bayesian learning framework of kernel machine and produces an identical functional form to the SVM. Tipping compared the RVM with SVM and demonstrated that the RVM has a comparable generalisation performance to the SVM and requires dramatically fewer kernel functions or model terms than the SVM. As Tipping stated, SVM suffers from its limitation of probabilistic prediction and Mercer's condition that it must be the continuous symmetric kernel of a positive integral operator. While RVM adopt a fully probabilistic framework and sparsity is achieved because the posterior distributions of many of the weights are sharply peaked around zero. The relevance vector comes from those training vectors associated with the remaining non-zero weights. However, a draw back of the RVM algorithm is a significant increase in computational complexity, compared with the SVM. Orthogonal least square (OLS) was first developed for the nonlinear data modelling, recently Chen [10–12] derived the locally regularised OLS (LROLS) algorithm to construct spares kernel models, which has shown to possess computational advantages compared with RVM. The LROLS only selects the significant terms, while RVM starts with the full model set. Moreover, LROLS only use a subset matrix of the full matrix that has been used by RVM. The subset matrix is diagonal and well-conditioned with small eigen-value spread. Further to Chen's research, Gao [13] has derived a critical vector learning (CVL) algorithm and improved the LROLS algorithm for the regression model, which has shown to possess more computational advantages. In this paper, the critical vector classification learning algorithm is applied to the text categorisation problem. Comparison results of SVM and CVL using the Reuters News Data Sets are presented and discussed.

The rest of this paper is organised as follows: Section 2 recalls the basic idea of SVM and explains its limitation compared with RVM. The algorithm of RVM with critical vector classification is presented in section 3. The detail implementation of applying critical learning algorithm in text categorisation is described in section 4. In section 5, the experiments are carried out using the Reuters data set, followed by the conclusions in section 6.

2 The Support Vector Machine

SVM is a learning system that uses a hypothesis space of linear functions in a high dimensional feature space. Joachims [8] explained the reason that SVM works well for text categorisation. Let's consider the binary classification problems about text document categorisation with SVM. Linear support vector machine trained on separable data. Let $f : X \subseteq R^n \rightarrow R$, where X is the vector set of documents. The input $x \in X$ is assigned to the positive class, if $f(x) \geq 0$; otherwise to negative class. When consider the $f(x)$ is a linear function, it can be rewritten as

$$f(x) = \langle w \cdot x \rangle + b = \sum_{i=1}^{n} w_i x_i + b \tag{1}$$

The basic idea of the support vector machine is to find the largest margin to do the classification in the hyper-plane, which means to minimise $\|w\|^2$, subject to

$$(x_i \cdot w) + b \geq +1 - \xi_i, \text{ for } y_i = +1, \tag{2}$$

$$(x_i \cdot w) + b \leq -1 + \xi_i, \text{ for } y_i = -1. \tag{3}$$

The optimal classification function is given by

$$g(x) = \text{sign}\{\langle w \cdot x \rangle + b\} \tag{4}$$

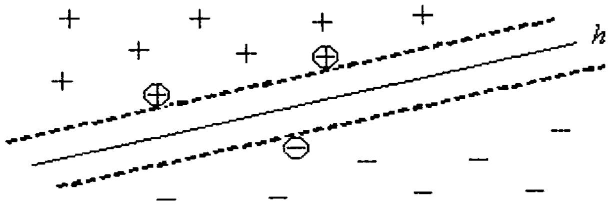

Fig. 1. Support vector machines find the hyper-plane h, which separates the positive and negative training examples with maximum margin. The examples closest to the hyper-plane in Figure 1 are called Support Vectors (marked with circles).

An appropriate inner product kernel $K(x_i, x_j)$ will be selected to realise the linear classification for non-linear problem. Then the function 1 can be written as:

$$y(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{N} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0 \tag{5}$$

Support vector machine has demonstrated successfully in many applications. However SVM suffers four major disadvantages: unnecessary use of basis functions; predictions are not probabilistic; entails a cross-validation procedure and the kernel function must satisfy Mercer's condition.

3 Critical Vector Learning

Tipping introduced the relevance vector machine (RVM), which does not suffer from the limitations mentioned in section 2. RVM can be viewed from a Bayesian learning framework of kernel machine and produces an identical functional form to the SVM. The functional form of the RVM is equivalent to the SVM. RVM generates predictive distributions which is a limitation of the SVM. And also RVM requires substantially fewer kernel functions.

Consider the scalar-valued target functions and giving the input-target pairs $\{\mathbf{x}_n, t_n\}_{n=1}^{N}$. The noise is assumed to be zero-mean Gaussian distribution with a variance of σ^2 . The likelihood of the complete data set can be written as

$$p\left(\mathbf{t}|\mathbf{w}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2\right\} \tag{6}$$

where $\mathbf{t} = (t_1 ... t_N)^T$, $\mathbf{w} = (w_0 ... w_N)^T$, and $\mathbf{\Phi} = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2) ..., \phi(\mathbf{x}_N)]^T$, wherein $\phi(\mathbf{x}_n) = [1, K(\mathbf{x}_n, \mathbf{x}_1), K(\mathbf{x}_n, \mathbf{x}_2), ..., K(\mathbf{x}_n, \mathbf{x}_N)]^T$. To make a simple function for the Gaussian prior distribution over w , 6 can be written as:

$$p(\mathbf{w}|\alpha) = \prod_{i=0}^{N} \mathcal{N}\left(w_i|0, \alpha_i^{-1}\right) \tag{7}$$

where α is a vector of $N+1$ hyper parameters.

Relevance vector learning can be looked as the search for the hyper parameter posterior mode, i.e. the maximisation of $p\left(\alpha, \sigma^2|\mathbf{t}\right) \propto p\left(\mathbf{t}|\alpha, \sigma^2\right) p\left(\alpha\right) p\left(\sigma^2\right)$ with respect to α and $\beta(\beta \equiv \sigma^2)$. RVM involves the maximisation of the product of the marginal likelihood and priors over α and σ^2. And MacKay [14] has given

$$\alpha_i^{new} = \frac{\gamma_i}{2\mu_i^2}, \ \beta^{new} = \frac{\|\mathbf{t} - \mathbf{\Phi}\mu\|^2}{N - \sum_i \gamma_i} \tag{8}$$

where μ_i is the $i-th$ posterior mean weight and N in the denominator refers to the number of data examples and not the number of basis functions. $\gamma_i \in [0,1]$ can be interpreted as a measure of how well-determined its corresponding parameter w_i is by the data.

A drawback of the RVM is a significant increase in computational complexity. Based on kernel methods and least squares algorithm, a locally regularised orthogonal least squares (LROLS) algorithm has been derived by Chen [10] to construct sparse kernel model.

$$y\left(k\right) = f\left(y\left(k-1\right), ..., y\left(k-n_y\right), u\left(k-1\right), ..., u\left(k-n_u\right)\right) + e\left(k\right)$$

$$y\left(k\right) = f\left(x\left(k\right)\right) + e\left(k\right) \tag{9}$$

where, $x\left(k\right) = \left[y\left(k-1\right), ..., y\left(k-n_y\right), u\left(k-1\right), ..., u\left(k-n_u\right)\right]^T$ denotes the system "input" vector, f is the unknown system mapping. Considering a general discrete-time nonlinear system represented by a nonlinear model, $u\left(k\right)$ and $y\left(k\right)$ are the system input and output variables, respectively, n_y and n_u are positive integers representing the lags in $y\left(k\right)$ and $u\left(k\right)$, respectively, $e\left(k\right)$ is the system white noise.

The system identification involves in construct a function (model) to approximate the unknown mapping f based on an N-sample observation dataset $D = \left\{x\left(k\right), y\left(k\right)\right\}_{k=1}^{N}$, i.e., the system input-output observation data $\left\{u\left(k\right), y\left(k\right)\right\}$. The most popular class of such approximating functions is the kernel regression model of the form:

$$y\left(k\right) = \widehat{y}\left(k\right) + e\left(k\right) = \sum_{i=1}^{N} \omega_i \phi_i\left(k\right) + e\left(k\right), \ 1 \leq k \leq N \tag{10}$$

where $\widehat{y}\left(k\right)$ denotes the "approximated" model output, ω_i's are the model weights, and $\phi_i\left(k\right) = k\left(x\left(i\right), x\left(k\right)\right)$ are the classifiers generated from a given kernel function $k\left(x, y\right)$ [15].

Focus on the single kernel function and by definitions in [13], the model can be viewed as the following matrix form:

$$y = \Phi\omega + e \tag{11}$$

The goal is to find the best linear combination of the columns of Φ (i.e. the best value for ω) to explain y according to some criterion. The normal criterion is to minimise the sum of squared errors,

$$E = e^T e \tag{12}$$

where the solution ω is called the least squares solution to the above model. Detail implementation is given in [16].

An equivalent regularisation formula can be adopted in the critical vector algorithm with PRESS statistic for the regularised objective [13]. The regularised critical vector algorithm with PRESS statistic is based on the following regularised error criterion

$$E(\omega, \alpha, \beta) = \beta e^T e + \sum_{i=1}^{n_M} \alpha_i \omega_i^2 = \beta e^T e + \omega^T H \omega \tag{13}$$

where n_M is the number of involved critical vectors, β is the noise parameter and $H = diag\left\{\alpha_1, ..., \alpha_{n_M}\right\}$ consisting of the hyper parameters used for regularising weights. The key issue in regularised regression formulation is to automatically optimise the regularisation parameter. The Bayesian evidence technique [14] can readily be used for this objective. Estimating hyper parameters is implemented in a loop procedure based on the calculation for α and β [17].

Define

$$A = \beta \Phi^T \Phi + H \tag{14}$$

and

$$\gamma_i = 1 - \alpha_i \left(A^{-1}\right)_{ii}, \quad \gamma = \sum_{i=1}^{n_M} \gamma_i \tag{15}$$

Then the update formulas for hyper parameters α_i and β can be given by

$$\alpha_i^{new} = \frac{\gamma_i}{2\omega_i^2}, \quad \beta^{new} = \frac{N - \gamma}{2e^T e} \tag{16}$$

The iterative hyper parameter and model selection procedure can be summarised:

Initialisation Set initial value for α_i and β for $i = 1, 2, ..., N$, for example, using estimated noise variance for the inverse of β and a small value 0.0001 for all α_i.

Step 1 Given the current α_i and β, use the procedure with PRESS statistic to select a subset model with critical vectors.

Step 2 Update α_i and β using equation 16. If α_i and β remains sufficiently unchanged in two successive iterations or a pre-set maximum iteration number is reached, then stop the algorithm; otherwise go to step 1.

4 Applying Critical Vector Learning in Text Categorisation

The document collection with n documents is represented by a term frequency document matrix

$$C = \begin{bmatrix} d_1 \\ \vdots \\ d_j \\ \vdots \\ d_n \end{bmatrix} \in \Re^{m \times n} \tag{17}$$

where document vector $d_j \in \Re^{m \times 1}$ represents the term frequency of m key terms in each document. The target variable

$$y = [y_1, \cdots, y_j, \cdots, y_n]^T \tag{18}$$

where y_j denotes the corresponding output of d_j, which represents the category that d_j belongs to.

The procedures of the training process was implemented as follows:

1. Calculate the keyword frequency of each document to construct the term frequency document matrix.
2. Construct the kernel matrix. Its (i, j)-th element is $K(d_i, d_j)$. Denote $\mathbf{x}_i$ as the i-th row of the kernel matrix Φ.
3. Select the k best $\mathbf{x}_i$ by repeating the following steps k times:
   (a) For every $\mathbf{x}_i$, use the least square algorithm to estimate the ω_i in equation 11.
   (b) Select the $\mathbf{x}_i$ with the smallest error.
   (c) Remove the i-th row of the kernel matrix (corresponding to the selected $\mathbf{x}_i$ f) to form a new matrix.
   (d) Remove the corresponding i-th element in the target variable vector y and form a new target variable as:

$$y = [y_1 - \mathbf{x}_i\omega_i, \cdots, y_{i-1} - \mathbf{x}_i\omega_i, y_{i+1} - \mathbf{x}_i\omega_i, \cdots, y_n - \mathbf{x}_i\omega_i]^T$$

4. Construct the training kernel model, $K\_training(\mathbf{x}_k) = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)$.

The prediction (or test) is conducted using the constructed training kernel.

5 Experimental Results

Experimental studies have been carried out to compare the performance of CVL and SVM. In this study, a java library for SVM (LIBSVM) was utilised while CVL was implemented using Scilab.

The Reuters News Data Sets, which has been frequently used as benchmarks for classification algorithms, has been used in this paper for the experiments. The

Reuters 21578 collection is a set of 21,578 short (average 200 words in length) news items, largely financially related, that have been pre-classified manually into 118 categories.

The experiments were conducted using 100 and 200 documents from three news group: C15 (performance group), C22 (new products/services group) and C21 (products/services group). The first set of experiments used C15 and C22 data, while the second set of experiments used C21 and C22. The second set of data is more difficult to classify than the first set since data sets C21 and C22 are closely related. This is confirmed by the experimental results, as shown in table 1 and table 2.

Table 1. Results of SVM and CVL classifiers on C15 and C22 data

No. of Documents	No. of Keywords	nSv (SVM)	Accuracy (SVM)	nSv (CVL)	Accuracy (CVL)
100	50	83	92.3%	13	91.02%
	100	83	92.3%	13	91.02%
200	50	122	92.4%	14	93.6%
	100	122	92.4%	14	93.6%

Table 2. Results of SVM and CVL classifiers on C21 and C22 data

No. of Documents	No. of Keywords	nSv (SVM)	Accuracy (SVM)	nSv (CVL)	Accuracy (CVL)
100	50	86	85.89%	14	84.61%
	100	86	85.89%	14	84.61%
200	50	153	84.81%	14	89.24%
	100	153	84.81%	14	89.24%

The result of the experiment shows that critical vector learning algorithm achieves the comparable accuracy with SVM. The advantage of using critical vector learning algorithm is that it requires dramatically fewer support vectors to construct the training model. This means it has less computation complexity and requires less computation time in conducting the prediction after the model is being built.

SVM performs slightly better when the number of document increase, while the CVL remain almost the same. However the number of support vectors required by SVM grows linearly with the size of the training set, while CVL various slightly.

The result of the experiment also shows that both SVM and CVL are not sensitive to the number of keywords, which the accuracy and the number of support vectors remain the same with different keyword attributes.

While SVM and CVL are implemented in different languages, comparison of computational time cannot be conducted at this stage. The next step is to implement CVL using JAVA which allows meaningful comparison of execution times.

6 Conclusions

The critical learning algorithm based on the kernel methods and least squares algorithm has achieves comparable classification accuracy to the SVM. SVM performs better when the number of document increase, but require much more support vectors with the size of the training set increasing. CVL requires slightly different number of the support vectors when the training set increase. The most benefit of CVL is that it requires dramatically fewer numbers of support vectors to construct the model. This will improve the prediction efficiency which is particularly useful in online applications.

References

1. Sebastiani, F.: Machine learning in automated text categorisation. ACM Computing Surveys **34** (2002) 1–47
2. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
3. Amasyali, M., Yildirim, T.: Automatic text categorization of news articles. In: Signal Processing and Communications Applications Conference, 2004. Proceedings of the IEEE 12th, Turkish (2004) 224 - 226
4. Basu, A., Walters, C., Shepherd, M.: Support vector machines for text categorization. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Hawaii (2003)
5. Hu, J., Huang, H.: An algorithm for text categorization with svm. In: IEEE Region 10 Conference on Computers, Communications,Control and Power Engineering. Volume 1., Beijin, China (2002) 47 - 50
6. Hu, X.Y.C., Chen, Y., Wang, L., Yun-Fa: Text categorization based on frequent patterns with term frequency. In: International Conference on Machine Learning and Cybernetics. Volume 3., Shanghai, China (2004) 1610 - 1615
7. Rennie, J.: Improving multi-class text classification with support vector machine (2001)
8. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: 10th European Conference on Machine Learning, Springer Verlag (1998) 137–142
9. Tipping, M.E.: Sparse bayesian learning and the relevance vector machine. Journal of Machine Learning Research **1** (2001) 211–244
10. Chen, S.: Locally regularised orthogonal least squares algorithm for the construction of sparse kernel regression models. In: 2002 6th International Conference on Signal Processing. Volume 2. (2002) 1229 - 1232
11. Chen, S., Hong, X., Harris, C.: Sparse kernel regression modeling using combined locally regularized orthogonal least squares and d-optimality experimental design. IEEE Transactions on Automatic Control **48** (2003) 1029 - 1036

12. Chen, S., Hong, X., Harris, C.: Sparse kernel density construction using orthogonal forward regression with leave-one-out test score and local regularization. IEEE Transactions on Systems, Man and Cybernetics, Part B **34** (2004) 1708 - 1717
13. Gao, J., Zhang, L., Shi, D.: Critical vector learning to construct sparse kernel modeling with press statistic. In: International Conference on Machine Learning and Cybernetics. Volume 5., Shanghai, China (2004) 3223 - 3228
14. MacKay, D.: Bayesian interpolation. IEEE Transactions on Neural Networks (1992) 415–447
15. Schlkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge, Massachusetts (2002)
16. Sun, P.: Sparse kernel least squares classifier. In: Fourth IEEE International Conference on Data Mining, Brighton, UK (2004) 539 – 542
17. Nabney, I.: Algorithms for Pattern Recognitions. Springer, London (2001)

AUSDM05

Proceedings
4th Australasian Data Mining Conference

5 - 6th December, 2005, Sydney, Australia

Edited by
Simeon J. Simoff, Graham J. Williams, John Galloway
and Inna Kolyshkina

Collocated with
the 18th Australian Joint Conference on
Artificial Intelligence AI2005
and
the 2nd Australian Conference on Artificial
Life ACAL05

University of Technology Sydney
2005

Proceedings of the 4th Australasian Data Mining Conference – AusDM05, 5 – 6th, December, 2005, Sydney, Australia, collocated with the 18th Australian Joint Conference on Artificial Intelligence AI2005 and the 2nd Australian Conference on Artificial Life ACAL05.

S. J. Simoff, G. J. Williams, J. Galloway and I. Kolyshkina (eds).

Conference Web Site:
http://www.togaware.com/ausdm05/

Published by the University of Technology Sydney

ISBN 1-86365-716-9

Supported by:

togaware	Web address: www.togaware.com
The e-Markets Research Group	Web address: www.e-markets.org.au
UNIVERSITY OF TECHNOLOGY SYDNEY INFORMATION TECHNOLOGY	Web address: www.uts.edu.au Web address: www.it.uts.edu.au
Institute of Analytics Professionals of Australia Limited	Web address: www.iapa.org.au
NetMap analytics	Web address: www.netmapanalytics.com
ARC Research Network on Data Mining and Knowledge Discovery	Web address: www.dmkd.flinders.edu.au

Foreword

The Australasian Data Mining Conference series **AusDM**, initiated in 2002, is the annual flagship venue where data mining and analytics professionals - scholars and practitioners, can present the state-of-art in the field. Together with the Institute of Analytics Professionals of Australia, **AusDM** has built a unique profile in nurturing this joint community. The first and second edition of the conference (held in 2002 and 2003 in Canberra, Australia) facilitated the links between different research groups in Australia and some industry practitioners. This year the event has been supported by:

- Togaware, again hosting the website and the conference management system, coordinating the review process and other essential expertise;
- the University of Technology, Sydney, providing the venue, registration facilities and various other support at the Faculty of Information Technology;
- the Institute of Analytic Professionals of Australia (IAPA) and NetMap Analytics Pty Limited, facilitating the contacts with the industry;
- the e-Markets Research Group, providing essential expertise for the event;
- the ARC Research Network on Data Mining and Knowledge Discovery, providing financial support.

The conference program committee reviewed 42 submissions, out of which 16 submissions have been selected for publication and presentation. **AusDM** follows a rigid blind peer-review process and ranking-based paper selection process. All papers were extensively reviewed by at least three referees drawn from the program committee. We would like to note that the cut-off threshold has been high (4.1 on a 5 point scale), which indicates the high quality of submissions. We would like to thank all those who submitted their work to the conference. We will be extending the conference format to be able to accommodate more presentations.

Today data mining and analytics technology has gone far beyond crunching databases of credit card usage or retail transaction records. This technology is a core part of the so-called "embedded intelligence" in science, business, health care, drug design, security and other areas of human endeavour. Unstructured text and richer multimedia data are becoming a major input to the data mining algorithms. Consistent and reliable methodologies are becoming critical to the success of data mining and analytics in industry. Accepted submissions have been grouped in four sessions reflecting these trends. Each session is preceded by invited industry presentation.

Special thanks go to the program committee members. The final quality of selected papers depends on their efforts. The **AusDM** review cycle runs on a very tight schedule and we would like to thank all reviewers for their commitment and professionalism.

Last but not least, we would like to thank the organisers of AI 2005 and ACAL 2005 for assisting in hosting **AusDM**.

Simeon, J. Simoff, Graham J. Williams
John Galloway and Inna Kolyshkina

November 2005

Conference Chairs

Simeon J Simoff	University of Technology, Sydney
Graham J Williams	Australian Taxation Office, Canberra
John Galloway	NetMap Analytics Pty Ltd, Sydney
Inna Kolyshkina	Pricewaterhouse Coopers Actuarial, Sydney

Program Committee

Hussein Abbass	University of New South Wales, ADFA, Australia
Helmut Berger	Electronic Commerce Competence Centre EC3, Austria
Jie Chen	CSIRO, Canberra, Australia
Peter Christen	Australian National University, Australia
Vladimir Estivill-Castro	Griffith University, Australia
Eibe Frank	University of Waikato, New Zealand
John Galloway	Netmap Analytics, Australia
Raj Gopalan	Curtin University, Australia
Mohamed Gaber	Monash University, Australia
Warwick Graco	Australian Taxation Office, Australia
Lifang Gu	CSIRO, Canberra, Australia
Simon Hawkins	University of Canberra, Australia
Robert Hilderman	University of Regina, Canada
Joshua Huang	Hong Kong University, China
Warren Jin	CSIRO, Canberra, Australia
Paul Kennedy	University of Technology, Sydney, Australia
Inna Kolyshkina	Pricewaterhouse Coopers Actuarial, Sydney, Australia
Jiuyong Li	University of Southern Queensland, Australia
John Maindonald	Australian National University, Australia
Arturas Mazeika	Free University Bolzano-Bozen, Italy
Mehmet Orgun	Macquarie University, Australia
Jon Patrick	The University of Sydney, Australia
Robert Pearson	Health Insurance Commission, Australia
Francois Poulet	ESIEA-Pole ECD, Laval, France
John Roddick	Flinders University
John Yearwood	University of Ballarat, Australia
Osmar Zaiane	University of Alberta, Canada

AusDM05 Conference Program, 5th – 6th December 2005, Sydney, Australia

Monday, 5 December, 2005

9:00 - 9:05 Opening and Welcome

09:05 - 10:05 INDUSTRY KEYNOTE "Text Mining"
Inna Kolyshkina, PricewaterhouseCoopers, Sydney

10:05 - 10:30 Coffee break

10:30 - 12:00 Session I: Text Mining

- 10:30 - 11:00 INCORPORATE DOMAIN KNOWLEDGE INTO SUPPORT VECTOR MACHINE TO CLASSIFY PRICE IMPACTS OF UNEXPECTED NEWS
  Ting Yu, Tony Jan, John Debenham and Simeon J. Simoff
- 11:00 - 11:30 TEXT MINING - A DISCRETE DYNAMICAL SYSTEM APPROACH USING THE RESONANCE MODEL
  Wenyuan Li, Kok-Leong Ong and Wee-Keong Ng
- 11:30 - 12:00 CRITICAL VECTOR LEARNING FOR TEXT CATEGORISATION
  Lei Zhang, **Debbie Zhang** and Simeon J. Simoff

12:00 - 12:30 DISCUSSION: Linking Analytics Industry and Academia
Convenor: Inna Kolyshkina, PricewaterhouseCoopers, Sydney

12:30 - 13:30 Lunch

13:30 - 14:30 INDUSTRY KEYNOTE "Network Data Mining"
John Galloway, NetMap Analytics, Sydney

14:30 - 15:00 Coffee break

15:00 - 17:00 Session II: Data Linking, Enrichment and Data Streams

- 15:00 - 15:30 ASSESSING DEDUPLICATION AND DATA LINKAGE QUALITY: WHAT TO MEASURE?
  Peter Christen and Karl Goiser
- 15:30 - 16:00 AUTOMATED PROBABILISTIC ADDRESS STANDARDISATION AND VERIFICATION
  Peter Christen and Daniel Belacic
- 16:00 - 16:30 DIFFERENTIAL CATEGORICAL DATA STREAM CLUSTERING
  Weijun Huang, Edward Omiecinski, Leo Mark and **Weiquan Zhao**
- 16:30 - 17:00 S-MONITORS: LOW-COST CHANGE DETECTION IN DATA STREAMS
  Weijun Huang, Edward Omiecinski, Leo Mark and **Weiquan Zhao**

Tuesday, 6 December, 2005

09:00 - 10:00 INDUSTRY KEYNOTE "The Analytics Profession: Lessons and Challenges"
Eugene Dubossarsky, Ernst & Young, Sydney

10:00 - 10:30 Coffee break

10:30 - 12:30 Session III: Methodological issues

- 10:30 - 11:00 DOMAIN-DRIVEN IN-DEPTH PATTERN DISCOVERY: A PRACTICAL METHODOLOGY
  Longbing Cao, Rick Schurmann, Chengqi Zhang
- 11:00 - 11:30 MODELING MICROARRAY DATASETS FOR EFFICIENT FEATURE SELECTION
  Chia Huey Ooi, Madhu Chetty, Shyh Wei Teng
- 11:30 - 12:00 PREDICTING INTRINSICALLY UNSTRUCTURED PROTEINS BASED ON AMINO ACID COMPOSITION
  Pengfei Han, Xiuzhen Zhang, Raymond S. Norton, and Zhiping Feng
- 12:00 - 12:30 A COMPARATIVE STUDY OF SEMI-NAIVE BAYES METHODS IN CLASSIFICATION LEARNING
  Fei Zheng and Geoffrey I. Webb

12:30 - 13:30 Lunch

13:30 - 14:30 INDUSTRY KEYNOTE "Analytics in The Australian Taxation Office"
Warwick Graco, Australian Taxation Office, Canberra

14:30 - 15:00 Coffee break

15:00 - 17:00 Session IV: Methodology and Applications

- 15:00 - 15:30 A STATISTICALLY SOUND ALTERNATIVE APPROACH TO MINING CONTRAST SETS
  Robert J. Hilderman and Terry Peckham
- 15:30 - 16:00 CLASSIFICATION OF MUSIC BASED ON MUSICAL INSTRUMENT TIMBRE
  Peter Somerville and Alexandra L. Uitdenbogerd
- 16:00 - 16:30 A COMPARISON OF SUPPORT VECTOR MACHINES AND SELF-ORGANIZING MAPS FOR E-MAIL CATEGORIZATION
  Helmut Berger and Dieter Merkl
- 16:30 - 17:00 WEIGHTED EVIDENCE ACCUMULATION CLUSTERING
  F. Jorge Duarte, Ana L. N. Fred, André Lourenço and M. Fátima C. Rodrigues

Table of Contents